Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: July 15, 1997

Period of Report: April 1, 1997 to June 30, 1997

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	7/14/97	Scientific/Tech

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D468

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR528181697

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS
Browsing, Query Processing, Indexing, Image Retrieval, Scanned Document Retrieval, Bayesian Network, Text Retrieval, Probabilistic Retrieval Model, Large Distributed Databases

15. NUMBER OF PAGES
8

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. Extensive use will be made of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query. We will also be making increased use of PTO text databases in these experiments.

Technical Results

We have developed a lexical acquisition program that can rapidly build up a vocabulary of phrases from a large corpus of text. An approach that combines a statistical approach and heuristics designed to avoid common mistakes was compared to an approach that relies on a part-of-speech tagger. The experiments demonstrated that the statistical approach was significantly faster and more accurate. We are currently building a default phrase vocabulary by analyzing a number of text corpora, including patents, that total more than 10 gigabytes. We will then start using the phrase vocabulary for patent indexing and retrieval experiments.

We have also recently downloaded the first 2.5 gigabytes of PTO Greenbook data. This was done over a slow internet link, and we anticipate more transfers after the faster link is installed next month. This patent data was indexed using the INQUERY system, which included representing all Greenbook fields. We also developed a demonstration of text search for this database on the web, including fielded queries, relevance feedback, short and long displays of patents, highlighting of fielded terms, and transforming the contents of some fields for more readable display. This demonstration is available at //darwin/cgi-bin/pto/pto_query/usr/dar/tmp4/pto/demo/demo. This is a temporary site and a link will be added to the U.Mass. PTO home page.

The work on "core concepts" has also been continued, with an emphasis on clustering the concepts in a complex query and using corpus characteristics to weight the resulting groups. This research is being done with the TREC corpus and may also be used as a test of the new probabilistic operators developed for the Bayesian net model.

Important Findings and Conclusions

The statistical lexical analysis approach is producing good results and is superior to an approach that uses more linguistic information in the form of part-of-speech tags. The probabilistic retrieval model appears to work well with basic patent queries.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

Now that we have built the initial PTO database, we can carry out more experiments with phrase representations, complex queries and core concepts in this environment.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff. Most of the classification experiments will be done in the context of the PTO classification and previously classified patents.

Technical Results

We continue to enhance the 3-D graphics interface for evaluation in the TREC interactive track this Fall. We have obtained the classification data and have developed classification software for large data sets, such as the PTO database.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

The availability of PTO data means that systematic classification experiments can now be carried out.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web..

Technical Results

We have continued to improve the indexing and retrieval techniques for images. A new indexing technique resulted in a fifty times speedup in query processing with very little loss in retrieval accuracy. A demonstration of this technique has been implemented for the web. We have recently downloaded 15 gigabytes of Patent image data which includes all the trademark images and text. Programs were written to convert both patent (Yellowbook) and Trademark (WIPO) images to standard TIFF format, and the trademark text was indexed with INQUERY. A subset of the trademark images (the latest 2,000) was used to create a test set for evaluating the appearance-based retrieval technique on the binary trademark data. The relatively poor quality of many of the images from PTO is a significant issue that may result in more experiments being done with external (e.g. web-based) databases.

Important Findings and Conclusions

Retrieval experiments using both appearance-based and new color matching techniques continue to produce very good accuracy. Indexing techniques show considerable promise for improving the efficiency of image retrieval.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

Experiments with appearance-based retrieval of trademarks have now started.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

More results on the distributed architecture have been obtained from simulations. Extending these simulations to be accurate for multi-terabyte databases is a major focus of our current work. We have also carried out more experiments with collection selection algorithms that have resulted in small improvements. The TREC Very Large Collection database has been downloaded and will be used for more experiments.

Important Findings and Conclusions

None.

Significant Hardware Development

None.

Special Comments

The fast network is expected to be available by the end of July.

Implications for Further Research

Scaling of both simulations and real implementations will continue.